Data Analysis and Visualization in R (IN2339)

Exercise Session 7 - Statistical Testing I

Felix Brechtmann, Jun Cheng, Vicente Yepez, Julien Gagneur

Section 00 - Getting Ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)
library(tidyr)
library(dplyr)
library(patchwork) # optional, makes plots nicer
```

2. Load the yeast data

```
genotype <- fread("./extdata/eqtl/genotype.txt")
genotype <- melt(genotype, id.vars = 'strain', variable.name = 'marker',
                 value.name = 'genotype')
growth <- fread("./extdata/eqtl/growth.txt")
growth <- melt(growth, id.vars = "strain", variable.name = 'media',
               value.name = 'growth_rate')
marker <- fread("./extdata/eqtl/marker.txt")
```

Section 01 - Permutation test of growth rate difference

1. The following code recreates the example shown in the lecture to test the association of the genotype at marker 5211 with the growth rate difference in Maltose medium. Note that the code is written using functions, meaning that it will work for any marker, not just marker 5211. Read it carefully to understand what happens in each function. Then execute the code.

```
# Plotting the growth rate difference
getMaltoseDt = function(mrk){
  growth_mrk <-  merge(growth, genotype[marker %in% mrk, .(strain, genotype, marker)],
                       by = 'strain', allow.cartesian = TRUE)
  growth_mrk[media == "YPMalt"]
}

# boxplot
plot_growth_one_mk <- function(mk){
    ggplot(getMaltoseDt(mk), aes(genotype, growth_rate)) +
    geom_boxplot() +
    labs(title = mk) + theme_bw(base_size = 16)
}
plot_growth_one_mk("mrk_5211")
```

```r
# Function to calculate the difference of the median of two genotypes
median_diff <- function(dt){
  dt[genotype == 'Wild isolate', median(growth_rate, na.rm=T)] -
    dt[genotype == 'Lab strain', median(growth_rate, na.rm=T)]
}

# Function to permute the table, plot the resulting histogram
# and compute a p-value
p_val_medians <- function(dt, N_permu = 1000){
  # It will return both a pvalue and plot a histogram of T_star
  T_ref <- median_diff(dt)
  T_star <- sapply(1:N_permu, function(x){
      median_diff(dt[, genotype := sample(genotype)]) })
  # Plot
  g <- ggplot(data = data.table(T_star = T_star), aes(T_star)) + geom_histogram() +
    geom_vline(aes(xintercept=T_ref, color="T_ref")) + xlim(-3,3)
  print(g) # Needed to render plot inside function call
  # Compute and return the p value
  p_val <- (sum(T_star > T_ref | T_star < -T_ref) + 1) / (N_permu + 1)
  p_val
}

# Calling the function:
p_val_medians(getMaltoseDt("mrk_5211"))
```

2. Using the code above, plot and test whether markers 1653 and 5091 associate with growth. Interpret your results.

Section 02 - Permutation test of marker association

1. We just concluded that both markers 5211 and 5091 are significantly associated with growth. However, this could be confounded. A common source of confounding in genomics is due to "linkage", which describes the phenomenon of markers being inherited together.

To investigate the issue of linkage in our dataset, test if marker 5091 significantly associates with marker 5211. Define a null hypothesis, a statistics and use permutation testing to answer the question. Strengthen your answer with a relevant plot.

Hint: start from:

```r
mks_geno <- genotype[marker %in% c('mrk_5091', 'mrk_5211')] %>%
  spread(marker, genotype)
```

and think about how this can be permuted.

Section 03 - Accounting for Confounding

1. We see that indeed marker 5211 and 5091 associate. Thus, the assocation between these markers and growth could be confounded.

We now would like to know if marker 5091 still associates with growth in maltose (YPMalt) when conditioned on marker 5211. Define a null hypothesis, a statistics and use permutation testing to answer the question. Strengthen your answer with a relevant plot.

2. Now, test if marker 5211 associates with growth in maltose when conditioned on marker 5091. Are the results the same? Discuss.

Section 04 - Confidence Intervals

1. Estimate 95% equi-tailed confidence intervals for the median of growth in maltose for each genotype at marker mrk_5211. Use the case resampling bootstrap scheme and report bootstrap percentile intervals. Propose a visualization of the results. Try it also with markers 5091 and 1653.